On the nature of
Phase-Type Poisson distributions

Sophie Hautphenne[*]   Guy Latouche[†]   Giang T. Nguyen[‡]

Abstract

Matrix-form Poisson probability distributions were recently introduced as one matrix generalization of Panjer distributions. We show in this paper that under the constraint that their representation is to be nonnegative, they have a physical interpretation as extensions of PH distributions, and we name this restricted family *Phase-type Poisson*. We use our physical interpretation to construct an EM algorithm-based estimation procedure.

1 Introduction

First appeared in Panjer (1981), Panjer's algorithm is designed to compute efficiently the density of sums of the form $S = \sum_{1 \leq i \leq N} X_i$, where the X_is are i.i.d. positive random variables and N is random, with a density $\{p_n\}$ that follows the recurrence relation

$$p_n = p_{n-1}(a + b/n) \qquad \text{for } n \geq 1, \tag{1}$$

p_0 being such that $\sum_{n \geq 0} p_n = 1$. If the X_is are nonnegative integer-valued random variables with density $\{f_n\}$, then the density $\{g_n\}$ of S may be

---

[*]The University of Melbourne, Department of Mathematics and Statistics, Victoria 3010, Australia; sophiemh@unimelb.edu.au

[†]Université Libre de Bruxelles, Département d'Informatique, CP 212, Boulevard du Triomphe, 1050 Bruxelles, Belgium; latouche@ulb.ac.be

[‡]Corresponding author: The University of Adelaide, School of Mathematical Sciences, SA 5005, Australia; giang.nguyen@adelaide.edu.au

recursively computed as

$$g_0 = p_0, \qquad g_n = \sum_{1 \le i \le n} f_i g_{n-i} (a + ib/n) \qquad \text{for } n \ge 1. \qquad (2)$$

This is a very efficient procedure, which has excellent numerical stability properties.

The distributions that satisfy (1) belong to a restricted set of families consisting of Poisson, binomial and negative binomial distributions (see Sundt and Jewell (1981)). Much effort has been spent to extend Panjer's algorithm to other distributions for N. In particular, its extension to Phase-type (PH) distributions is of great interest: since they are dense in the class of distributions on $\mathbb{N}$, this significantly increases the applicability of Panjer's algorithm.

Phase-type distributions have been introduced by Neuts (1975) and (1981) and they may be defined algebraically as follows: consider a sub-stochastic matrix T of order m such that $I - T$ is nonsingular, a density vector $\boldsymbol{\alpha}$ of order m, and define a sequence $\{\boldsymbol{v}_n\}$ of row vectors with

$$\boldsymbol{v}_1 = \boldsymbol{\alpha}(I - T), \qquad \boldsymbol{v}_n = \boldsymbol{v}_{n-1} T \qquad \text{for } n \ge 2. \qquad (3)$$

The density $p_0 = 1 - \boldsymbol{\alpha}\mathbf{1}$, $p_n = \boldsymbol{v}_n \mathbf{1}$, for $n \ge 1$, where $\mathbf{1}$ is a column vector of ones, is said to be of phase-type, with representation $(\boldsymbol{\alpha}, T)$. There is a clear similarity between (1) and (3), which suggests that the recursion (2) might be adapted to provide an efficient and numerically stable algorithm to compute the density of S when N has a PH distribution. This is done in two recent papers, Wu and Li (2010) and Siaw et al. (2011). The former defines the *generalized $(a, b, 0)$ family* as

$$p_n = \boldsymbol{\gamma} P_n \mathbf{1} \qquad \text{for } n \ge 0, \qquad (4)$$

where the matrices $\{P_n\}$ of order m are recursively defined as follows:

$$P_n = P_{n-1}(A + \frac{1}{n}B) \qquad \text{for } n \ge 1. \qquad (5)$$

The parameters are the matrices A, B, P_0 and the vector $\boldsymbol{\gamma}$, which is assumed to be nonnegative and normalized, so that $\boldsymbol{\gamma}\mathbf{1} = 1$.

Siaw et al. (2011) define the *generalized $(a, b, 1)$ family*, the difference being that the recursion (5) starts at $n = 2$, and the parameters are A, B, P_1 and p_0, while the matrix P_0 becomes irrelevant. The PH$(\boldsymbol{\alpha}, T)$ distribution belongs to the generalized $(a, b, 1)$ family, with $A = T$, $B = 0$, $p_0 = 1 - \boldsymbol{\alpha}\mathbf{1}$, $\boldsymbol{\gamma} = (\boldsymbol{\alpha}\mathbf{1})^{-1}\boldsymbol{\alpha}$ and $P_1 = (\boldsymbol{\alpha}\mathbf{1})(I - T)$.

The core of the algorithm in Wu and Li (2010) and Siaw et al. (2011) is the vector recursion

$$\boldsymbol{h}_n = \sum_{1 \leq i \leq n} f_i \boldsymbol{h}_{n-i} (A + \frac{1}{i} B) \tag{6}$$

to replace (2), with $g_n = \boldsymbol{h}_n \boldsymbol{1}$. Ren (2010) gives an improved algorithm in case N and the X_is themselves are of phase-type. Finally, we note that PH distributions have rational generating functions, and this is the basis for the adaptation in Eisele (2006) of Panjer's algorithm to the case where N is PH. A comparison of the complexity and numerical stability of the algorithms in Eisele (2006), Ren (2010), Wu and Li (2010) and Siaw et al. (2011) is outside the scope of the present paper.

We expect the generalized $(a, b, 0)$ and $(a, b, 1)$ distributions to form a very rich family since they include the PH distributions. However, as we show in the next section, the combination of two matrices in (5) makes these distributions a bit unwieldy, unless one imposes some simplifying constraint. In Section 2, we show that the series $\sum_{n \geq 0} P_n$ is a key quantity and that, for all practical purpose, it is necessary that the spectral radius of A be strictly less than one in order for the series to converge. Before doing so, we briefly address the issue of the choice of representation, and we adopt one that is slightly different from the representation in Wu and Li (2010) and Siaw et al. (2011).

Next, we assume in Section 3 that A and B commute. As matrices go, this is a very strong constraint, but it considerably simplifies the determination of the generating function and of moments, and it is a property of all the examples in Wu and Li (2010) and Siaw et al. (2011). In Section 4, we focus our attention on distributions for which $A = 0$, $B \geq 0$, and $\boldsymbol{\gamma} \geq \boldsymbol{0}$. These distributions are interesting because they form a family totally distinct from PH distributions, yet they are amenable to a Markovian representation. For that reason, we call them *Phase-type Poisson* or PH-Poisson distributions. This physical interpretation opens the way in Section 5 to an estimation procedure based on the EM algorithm.

2 Matrix generating function

We are concerned with distributions $\{p_n\}$ defined as

$$p_n = \boldsymbol{\beta} P_n \boldsymbol{1}, \qquad \text{where} \qquad P_n = \prod_{1 \leq i \leq n} (A + \frac{1}{i} B) \qquad \text{for } n \geq 0, \tag{7}$$

3

A and B are matrices of order m, and $\boldsymbol{\beta}$ is a row vector of size m. We use the convention that for $n = 0$, the matrix product in (7) is equal to the identity matrix, so that we may recursively define the P_ns as

$$P_0 = I, \qquad P_n = P_{n-1}(A + \frac{1}{n}B) \qquad \text{for } n \geq 1. \tag{8}$$

We shall write that $\{p_n\}$ has the representation $\mathcal{D}(\boldsymbol{\beta}, A, B)$ of order m.

This definition calls for a few comments. First, we assume that the recursion (7) starts with $n = 0$. In other words, we are not concerned in this paper with the possibility that the sequence $\{p_n\}$ does not conform to the general pattern for small values of n. Instead, we focus our attention, to a large extent, on the matrices P_n.

Second, our definition is slightly different from that of generalized $(a, b, 0)$ distributions in Wu and Li (2010), where it is assumed that $\boldsymbol{\gamma}$ is a stochastic vector ($\boldsymbol{\gamma} \geq \mathbf{0}$, $\boldsymbol{\gamma}\mathbf{1} = 1$) and that P_0 is a matrix chosen according to the circumstances. The two representations are equivalent as it suffices to define $\boldsymbol{\beta} = \boldsymbol{\gamma}P_0$. Our reason to prefer (7) is that we do not find any advantage in requiring that $\boldsymbol{\beta}$ should be stochastic when A, B and P_0 are allowed to be of mixed signs. Furthermore, our definition involves m^2 fewer parameters (the entries of P_0) and this savings will prove significant in Section 5 when we design an estimation procedure.

Finally, one might use left- instead of right-multiplication and define $P_n = (A + \frac{1}{n}B)P_{n-1}$, yielding a possibly different family of distributions. Actually, we shall assume in the next section that A and B commute, so that there would be no difference.

We need to impose some constraints on the representations of these distributions, otherwise very little can be said in general. To begin with, let us associate a transition graph to the matrices A and B: the graph contains m nodes, and there is an oriented arc from i to j if $|A_{ij}| + |B_{ij}| \neq 0$. A node j is said to be useful if there exists a node i such that there is a path from i to j in the transition graph *and* such that $\beta_i \neq 0$; j is said to be useless otherwise. The lemma below shows that one may require without loss of generality that representations are chosen without useless nodes.

Lemma 2.1 *If the representation $\mathcal{D}(\boldsymbol{\beta}, A, B)$ of order m is such that there exists at least one useless node, then there exists another, equivalent, representation of order m' strictly less than m.*

Proof Assume that j is a useless node; define S_1 to be the subset of nodes containing j and all the nodes i for which there exists a path from i to j.

4

The matrices A and B may be written, possibly after a permutation of rows and columns, as

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{1,1} & B_{1,2} \\ 0 & B_{2,2} \end{bmatrix},$$

where $A_{1,1}$ and $B_{1,1}$ are indexed by the nodes in S_1 and $A_{2,2}$ and $B_{2,2}$ are indexed by the remaining nodes; similarly, we have

$$P_n = \begin{bmatrix} (P_n)_{1,1} & (P_n)_{1,2} \\ 0 & (P_n)_{2,2} \end{bmatrix}$$

with $(P_n)_{1,1} = \prod_{1 \leq i \leq n}(A_{1,1} + \frac{1}{i}B_{1,1})$ and $(P_n)_{2,2} = \prod_{1 \leq i \leq n}(A_{2,2} + \frac{1}{i}B_{2,2})$.

We partition $\boldsymbol{\beta}$ in a similar manner and write $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \end{bmatrix}$. Since j is useless, $\boldsymbol{\beta}_1 = \mathbf{0}$. It is clear that $\boldsymbol{\beta}P_n\mathbf{1} = \boldsymbol{\beta}_2(P_n)_{2,2}\mathbf{1}$, so that $\mathcal{D}(\boldsymbol{\beta}_2, A_{2,2}, B_{2,2})$ is an equivalent representation, of order strictly smaller than m. $\qquad\square$

The generating function $p(z) = \sum_{n \geq 0} z^n p_n$ may be written as $p(z) = \boldsymbol{\beta}P(z; A, B)\mathbf{1}$, where

$$P(z; A, B) = \sum_{n \geq 0} z^n P_n = \sum_{n \geq 0} z^n \prod_{1 \leq i \leq n}(A + \frac{1}{i}B), \tag{9}$$

provided that the series in (9) converges. We focus our attention on the matrix generating function $P(z; A, B)$ and we discuss its convergence properties as $z \to 1$. A simple condition for $P(1; A, B)$ to be finite is given in the next lemma.

Lemma 2.2 *If* $\mathrm{sp}(A) < 1$, *where* $\mathrm{sp}(\cdot)$ *denotes the spectral radius, then the series* $P(z; A, B)$ *converges for* $|z| \leq 1$.

If $A \geq 0$ *and* $B \geq 0$, *then the inequality* $\mathrm{sp}(A) < 1$ *is both necessary and sufficient.*

Proof The convergence radius R of the series in (9) is given by $R^{-1} = \limsup_n \sqrt[n]{\|P_n\|}$, where $\|\cdot\|$ is any matrix norm. To simplify the notations, we define $C_i = A + (1/i)B$. For any consistent norm, $\|P_n\| \leq \|C_1\|\|C_2\| \cdots \|C_n\|$. Furthermore, $\|C_i\| \leq \|A\| + (1/i)\|B\|$ and for any $\varepsilon > 0$, there exists a norm such that $\|A\| < \mathrm{sp}(A) + \varepsilon$.

This implies that if $\mathrm{sp}(A) < 1$, then there exist $\eta < 1$ and i^* such that $\|C_i\| < \eta$ for all $i \geq i^*$. In addition,

$$\|P_n\|^{1/n} \leq \|C_1 C_2 \cdots C_{i^*}\|^{1/n}(\|C_{i^*+1}\| \cdots \|C_n\|)^{1/n}$$
$$\leq \|C_1 C_2 \cdots C_{i^*}\|^{1/n}\eta^{(n-i^*)/n},$$

for $n \geq i^*$, and $\|C_1 C_2 \cdots C_{i^*}\|^{1/n} \eta^{(n-i^*)/n} \to \eta$ as $n \to \infty$. We conclude, therefore, that $\limsup_n \sqrt[n]{\|P_n\|} \leq \eta$ and $R \geq 1/\eta > 1$, which proves the first claim.

If A and B are non-negative, then $P_n \geq A^n$ and $P(1; A, B) \geq \sum_{n \geq 0} A^n$; since the last series diverges if $\mathrm{sp}(A) \geq 1$, this completes the proof of the second claim. $\qquad\square$

Note that $\mathrm{sp}(A) < 1$ cannot be a necessary condition in all generality: to give one example, if there is some n^* such that $P_n = 0$ for all $n > n^*$, then the series in (9) reduces to a finite sum, and the spectral radius of A has no bearing on its convergence; such is the case if $B = -n^* A$.

We turn our attention to the derivatives

$$M_n(A, B) = \left. \frac{\partial^n}{\partial z^n} P(z; A, B) \right|_{z=1}, \tag{10}$$

for $n \geq 1$, assuming that they exist. In that case, the factorial moments of the distribution are given by $m_n = \boldsymbol{\beta} M_n(A, B) \mathbf{1}$. From the proof of Lemma 2.2, if $\mathrm{sp}(A) < 1$, then $P(z; A, B)$ is a matrix of analytic functions in the closed unit disk, and it is a sufficient condition for the derivatives to be finite at $z = 1$.

Lemma 2.3 *The matrices $M_n(A, B)$ are given by*

$$M_n(A, B) = n! P_n P(1; A, nA + B). \tag{11}$$

If $\mathrm{sp}(A) < 1$*, then we also have*

$$M_n(A, B) = n! P(1; A, B) \widetilde{P}_n, \tag{12}$$

where

$$\widetilde{P}_n = \prod_{1 \leq i \leq n} ((A + \frac{1}{i}B)(I - A)^{-1}).$$

Proof We write

$$P_n = \frac{1}{n!} \prod_{1 \leq i \leq n} (iA + B) = \frac{1}{n!}(A + B) \prod_{1 \leq i \leq n-1} (iA + (A + B))$$

$$= \frac{1}{n}(A + B) \prod_{1 \leq i \leq n-1} (A + \frac{1}{i}(A + B))$$

so that

$$\frac{\partial}{\partial z} P(z; A, B) = \sum_{n \geq 1} n z^{n-1} P_n = (A + B) \sum_{n \geq 1} z^{n-1} \prod_{1 \leq i \leq n-1} (A + \frac{1}{i}(A + B))$$

$$= (A + B) P(z; A, A + B)$$

6

and, by induction,

$$\frac{\partial^n}{\partial z^n}P(z;A,B) = (A+B)(2A+B)\cdots(nA+B)P(z;A,nA+B)$$
$$= n!P_nP(z;A,nA+B)$$

for all n, from which (11) results.

On the other hand, Lemma 1 in Wu and Li (2010) states that

$$\frac{\partial}{\partial z}P(z;A,B) = z\frac{\partial}{\partial z}P(z;A,B)A + P(z;A,B)(A+B).$$

If $\mathrm{sp}(A) < 1$, then

$$\frac{\partial}{\partial z}P(z;A,B) = P(z;A,B)(A+B)(I-zA)^{-1}$$

for $|z| \leq 1$, from which (12) readily results by induction. $\qquad\square$

This lemma points to the importance of being able to determine the matrix $P(1;A,B)$. In some special cases, an explicit expression may be derived but in general, in the absence of any simplifying feature of the pair (A,B), there does not seem to be an alternative to the brute force calculation of the series $\sum_{n\geq 0}\prod_{1\leq i\leq n-1}(A+\frac{1}{i}B)$.

3 Commutative matrix product

In this section, we assume that A and B commute and thereby obtain a stronger result than in Section 2. This assumption is satisfied for all examples in Wu and Li (2010), where either $A = 0$ or B is a scalar multiple of A. It is also satisfied if A or B is a scalar matrix cI for some scalar c, or if $B = 0$. The latter includes $\mathrm{PH}(\boldsymbol{\alpha},T)$ distributions if there exists a solution to the system of linear constraints $\boldsymbol{\alpha}(I-T)T^{n-1}\mathbf{1} = \boldsymbol{\beta}A^n\mathbf{1}$ for $n \geq 1$; if T is invertible, then an obvious solution is $\boldsymbol{\beta} = \boldsymbol{\alpha}(I-T)T^{-1}$, $A = T$.

Thus, although it is a restrictive assumption from a linear algebraic point of view, it may be reasonable in the context of stochastic modeling.

Theorem 3.1 *If A and B commute, then $P(z;A,B) = e^{(A+B)D(z;A)}$, where*

$$D(z;A) = z\sum_{n\geq 1}\frac{1}{n}(zA)^{n-1}$$
$$= A^{-1}\log(I-zA)^{-1} \qquad \textit{if } A \textit{ is nonsingular.}$$

Furthermore, if $B = -kA$ for some integer $k \geq 1$, then $P(z;A,B) = (I-zA)^{k-1}$ and $P(1;A,B)$ is finite, otherwise, $P(1;A,B)$ converges if and only if $\mathrm{sp}(A) < 1$.

Proof First, we observe that

$$\prod_{1 \le i \le n} (iA + B) = \sum_{0 \le i \le n} \begin{bmatrix} n \\ i \end{bmatrix} (A + B)^i A^{n-i} \tag{13}$$

where $\begin{bmatrix} n \\ i \end{bmatrix}$ are Stirling's numbers of the first kind. If A and B are scalars, then (13) is a straightforward consequence of the definition of Stirling's numbers in Knuth (1968), Section 1.2.6, equation (40). To prove the extension to commuting matrices, one proceeds by induction, using

$$\begin{bmatrix} n \\ 0 \end{bmatrix} = 0, \qquad \begin{bmatrix} n \\ n \end{bmatrix} = 1, \qquad \text{and} \qquad \begin{bmatrix} n \\ i-1 \end{bmatrix} + n \begin{bmatrix} n \\ i \end{bmatrix} = \begin{bmatrix} n+1 \\ i \end{bmatrix}$$

for $n \ge 1$. Next, we write

$$P(z; A, B) = \sum_{k \ge 0} z^k \frac{1}{k!} \sum_{0 \le i \le k} \begin{bmatrix} k \\ i \end{bmatrix} (A + B)^i A^{k-i} \qquad \text{by (9, 13)} \tag{14}$$

$$= \sum_{i \ge 0} z^i (A + B)^i \sum_{k \ge i} \frac{1}{k!} \begin{bmatrix} k \\ i \end{bmatrix} (zA)^{k-i} \tag{15}$$

since, as we show later, we may interchange the order of summation. By equations (25) and (26) in Knuth (1968), Section 1.2.9,

$$i! \sum_{k \ge i} \frac{1}{k!} \begin{bmatrix} k \\ i \end{bmatrix} x^{k-i} = (\sum_{k \ge 1} \frac{1}{k} x^{k-1})^i = (x^{-1} \log(1 - x)^{-1})^i,$$

and (15) becomes

$$P(z; A, B) = \sum_{i \ge 0} \frac{1}{i!} ((A + B)D(z; A))^i,$$

which proves the first claim.

If $B = -kA$, then

$$P(z; A, B) = e^{(1-k)zA \sum_{n \ge 1} (zA)^{n-1}/n} = e^{(k-1)\log(I - zA)} = (I - zA)^{k-1}.$$

Thus, it remains for us to justify the transition from (14) to (15). To that end, we show that the series is absolutely convergent if and only if $\mathrm{sp}(A) < 1$. It is well-known that $\|A^k\| = O(1)\mathrm{sp}(A)^k k^r$ asymptotically as $k \to \infty$, for some integer $r \ge 0$. Furthermore, $\begin{bmatrix} k \\ i \end{bmatrix}/k! = O(1)(\log k)^{i-1}/(i-1)!$, by Theorem 1 in Wilf (1993). Therefore,

$$\lim_{k \to \infty} \sqrt[k]{\frac{1}{k!} \begin{bmatrix} k \\ i \end{bmatrix} \|A^{k-i}\|} = \mathrm{sp}(A) \lim_{k \to \infty} \sqrt[k]{k^r (\log k)^{i-1}} = \mathrm{sp}(A)$$

8

so that the series $\sum_{k \geq i} \frac{1}{k!} \begin{bmatrix} k \\ i \end{bmatrix} (zA)^{k-i}$ in (15) absolutely converges in $\|z\| \leq 1$ if and only if $\mathrm{sp}(A) < 1$, in which case its limit is $\frac{1}{i!}(D(z;A)/z)^i$. The equation (15) becomes

$$P(z; A, B) = \sum_{i \geq 0} \frac{1}{i!}((A + B)D(z;A))^i$$

which converges without further constraint. □

This theorem confirms the important role of the matrix A with respect to the convergence of various series. A direct consequence is that if A, B_1 and B_2 are three commuting matrices, then

$$P(z; A, B_1)P(z; A, B_2) = e^{(A+B_1)D(z;A)}e^{(A+B_2)D(z;A)} = e^{(2A+B_1+B_2)D(z;A)}$$
$$= P(z; A, A + B_1 + B_2), \qquad (16)$$

so that, if A and B commute, we may write that

$$P(z; A, kA + B) = P(z; A, B)(P(z; A, 0))^k = P(z; A, B)(e^{AD(z;A)})^k$$
$$= P(z; A, B)(I - zA)^{-k}$$

for $k \geq 0$, k integer, and we may state the following property, using either (11) or (12):

Corollary 3.2 *If A and B commute, then the nth factorial moment of the distribution is given by*

$$m_n(\boldsymbol{\beta}, A, B) = n!\boldsymbol{\beta}P(1; A, B)(I - A)^{-n}P_n\mathbf{1}. \qquad (17)$$

□

If one remembers that $\boldsymbol{\beta}P(1; A, B)$ is a vector of which the components add-up to one, the similarity with the factorial moments of discrete PH distributions is striking (see equation (2.15) of Latouche and Ramaswami (1999)).

To conclude this section, we review the examples in Wu and Li (2010):

- If $B = \alpha A$ for $\alpha \geq -1$, then $(A + B)D(z; A) = (1 + \alpha) \log(I - zA)^{-1}$ and $P(z; A, \alpha A) = (I - zA)^{-(1+\alpha)}$.

- If $B = -kA$ for $k \geq 0$, k integer, $P(z; A, -kA) = (I - zA)^{k-1}$ as proved in Theorem 3.1.

- If $A = 0$, then $D(z; 0) = z$ and $P(z; 0, B) = e^{zB}$.

We shall further examine this last case in the remainder of the paper.

9

4 PH-Poisson distributions

4.1 Definition and comparison to PH distributions

We restrict our attention to distributions for which $A = 0$, with the added constraint that $\boldsymbol{\beta} \geq \mathbf{0}$, $B \geq 0$. The assumption that $\boldsymbol{\beta}$ and B are non-negative makes it easier to ascertain that $\mathcal{D}(\boldsymbol{\beta}, 0, B)$ is the representation of a probability distribution. In addition, as we show in Theorem 4.4, it provides us with a physical interpretation in terms of a Markovian process, which explains why we call these *Phase-type Poisson* distributions, or PH-Poisson for short.

Definition 4.1 *A random variable X has a PH-Poisson distribution with representation $\mathcal{P}(\boldsymbol{\beta}, B)$ if*

$$P[X = n] = p_n = \frac{1}{n!} \boldsymbol{\beta} B^n \mathbf{1}, \qquad for \ n \geq 0, \tag{18}$$

where $B \geq 0$ is a matrix of order m and $\boldsymbol{\beta} \geq \mathbf{0}$ is a row-vector of size m such that $\boldsymbol{\beta} e^B \mathbf{1} = 1$.

Note that $\boldsymbol{\beta} \mathbf{1} < 1$, unless $B = 0$. In the notation of Wu and Li (2010), the PH-Poisson distribution with representation $\mathcal{P}(\boldsymbol{\beta}, B)$ belongs to the generalized $(a, b, 0)$ family, with $A = 0$, $B = B$, $P_0 = e^{-B}$ and $\boldsymbol{\gamma} = \boldsymbol{\beta} e^B$. The generating function $p(z) = \sum_{n \geq 0} z^n p_n$ is given by $p(z) = \boldsymbol{\beta} e^{zB} \mathbf{1}$ and the factorial moments by

$$E[X(X-1) \cdots (X-n+1)] = \boldsymbol{\beta} B^n e^B \mathbf{1}. \tag{19}$$

It is easy to see that PH and PH-Poisson distributions are essentially two different families of probability distributions. Indeed, assume that X is PH-Poisson with representation $\mathcal{P}(\boldsymbol{\beta}, B)$ and Y is PH with representation $\text{PH}(\boldsymbol{\alpha}, T)$. From (18), it results that

$$P[X = n] \approx (\text{sp}(B))^n n^r / n!, \tag{20}$$

asymptotically as $n \to \infty$, where r is the index of $\text{sp}(B)$, and similarly,

$$P[Y = n] = \boldsymbol{\alpha} T^{n-1} (I - T) \mathbf{1} \approx (\text{sp}(T))^n n^s, \tag{21}$$

where s is the index of $\text{sp}(T)$. It is obvious that for any given B there is no T such that the right-hand sides of (20) and (21) coincide for all n big enough, unless $\text{sp}(B) = \text{sp}(T) = 0$.

If $\text{sp}(B) = 0$, then there exists $k \leq m$ such that $B^k = 0$, the distribution of X is concentrated on $\{0, 1, \ldots, k\}$, and X does have a PH representation by Theorem 2.6.5 in Latouche and Ramaswami (1999).
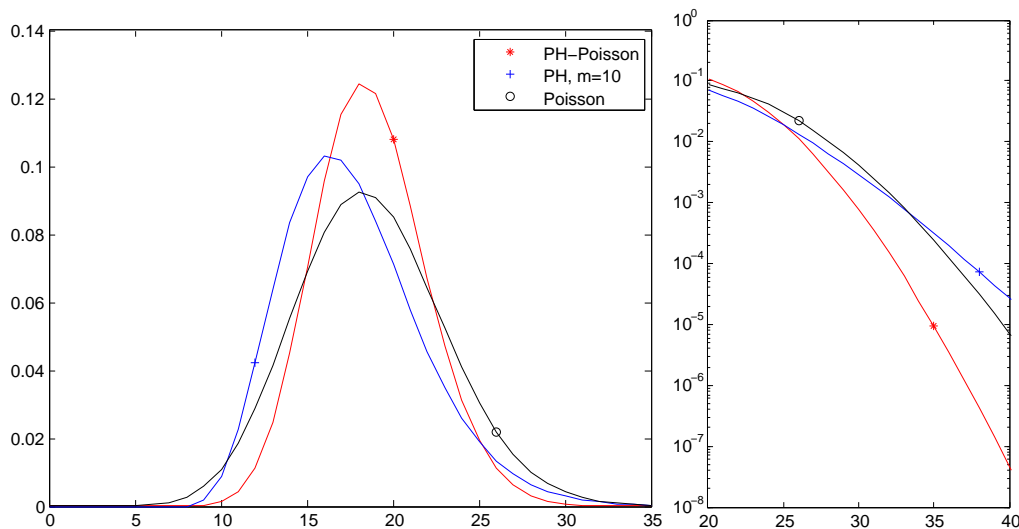
Figure 1: Density function for a PH-Poisson distribution with $m = 10$ and mean 18.71 (curve marked with $*$), the Poisson distribution with the same mean (marked with $+$) and the minimal-variance PH distribution with the same order and the same mean (marked with $\circ$).

Example 4.2 Tail of the density. We see from (20) that the density of PH-Poisson distributions drops sharply to zero. We compare three different densities on Figure 1: one is a PH-Poisson distribution, the second a PH distribution and the third a Poisson distribution. We have connected the points of the densities for better visual appearance, and we plot on the right-hand side the tail of the densities in semi-logarithmic scale.

The curve marked with a "$*$" is the density of the PH-Poisson distribution with $m = 10$, $B_{ii} = 10$, $1 \leq i \leq m$, and $B_{i,i+1} = 37.5$, $1 \leq i \leq m - 1$. The vector $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \left(\mathrm{diag}(e^B \mathbf{1})\right)^{-1}$$

(it is easy to verify that $\boldsymbol{\beta} e^B \mathbf{1} = 1$.) Its mean μ, variance σ^2 and coefficient of variation C.V. equal to σ/μ are given in the first row of Table 1, as well as the spectral radius S.R. of the matrix B.

The curve marked with a "$+$" is the density of the PH distributions with the same order m and mean μ and minimal variance (see Telek (2000) for details). The curve marked with a "$\circ$" is the Poisson density with parameter equal to the mean μ. The variance, coefficient of variation, and spectral radius of these two densities are also given in Table 1.

The plot on the right-hand side of Figure 1 clearly indicates that the PH-Poisson density decays asymptotically the fastest of the three, this is due to

11

	μ	σ^2	C.V.	S.R.
PH-Poisson	18.71	10.35	0.17	10
Phase-type	"	16.30	0.22	0.47
Poisson	"	18.71	0.23	18.71

Table 1: Mean, variance and coefficient of variation of the three distributions of Figure 1.

	μ	σ^2	C.V.	S.R.
PH-Poisson	37.71	73.89	1.96	42
Phase-type, $m = 10$	"	104.5	2.77	0.73
Phase-type, $m = 13$	"	71.69	1.90	0.66

Table 2: Mean, variance and coefficient of variation of the three distributions of Figure 2.

the combination of a relatively small spectral radius and of the factor $1/n!$. We also see from the plot on the left-hand side, and from Table 1, that it is the most concentrated around the mean.

Example 4.3 Small variance. We pursue here the comparison between PH-Poisson distributions and minimal variance PH distributions, showing that PH-Poisson distributions may prove to be a useful alternative to PH distributions when modeling discrete distributions with small variance.

The curve marked with a "$*$" on Figure 2 is the density of the PH-Poisson distribution with $m = 10$, $B_{ii} = 2 + 4i$, $1 \leq i \leq m$, and $B_{i,i+1} = 0.5$, $1 \leq i \leq m - 1$. The vector $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta} = \frac{1}{m} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} (\mathrm{diag}(e^B \mathbf{1}))^{-1}.$$

Its mean, variance, coefficient of variation and spectral radius are given in Table 2.

The two other curves are the density functions of PH distributions with minimal variance, with the same mean as the PH-Poisson distributions, and with different orders. The one marked with "$+$" has the same order $m = 10$ as the PH-Poisson distribution, the one marked with "$\circ$" has order $m = 13$, the smallest value for which the minimal variance is smaller than that of the PH-Poisson distribution.

4.2   A physical interpretation

We now give a physical interpretation for PH-Poisson distributions. First, we define the Poisson process $\{\theta_1, \theta_2, \ldots\}$ of rate $\nu = \max_i (B\mathbf{1})_i$. Second, we
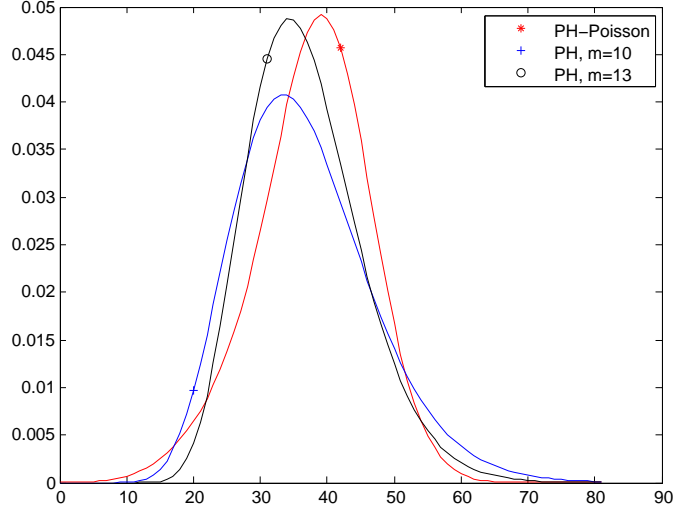
Figure 2: Density function for a PH-Poisson distribution with $m = 5$ (curve marked with $*$), a minimal-variance PH distribution with $m = 5$ (marked with $+$) and a minimal-variance PH distribution with $m = 17$ (marked with $\circ$). The distributions all have the same mean $\mu = 34.00$.

define $P = \nu^{-1}B$; P is a sub-stochastic matrix, possibly stochastic. Next, we consider a discrete PH random variable K with representation $(\boldsymbol{\alpha}, P)$, where $\boldsymbol{\alpha} = c\boldsymbol{\beta}$ for some arbitrary but fixed constant $c \leq (\boldsymbol{\beta}\mathbf{1})^{-1}$. In the present description, the Markov chain with transition matrix P makes a transition at each event of the Poisson process and it gets absorbed at time $T = \theta_K$. Finally, we count the number $N(t)$ of transitions between *transient* states until the Markov chain enters its absorbing state; that is, $N(t)$ is the number of Poisson events in the interval $(0, t)$ for $t < T$ and $N(t) = K - 1$ for $t \geq T$.

Theorem 4.4 *If $\boldsymbol{\alpha} = c\boldsymbol{\beta}$ for some arbitrary but fixed constant $c \leq (\boldsymbol{\beta}\mathbf{1})^{-1}$, then p_n defined in (18) is the conditional probability*

$$p_n = \mathrm{P}[N(1) = n | T > 1]. \tag{22}$$

Proof Define $M_k(t)$ such that

$$(M_k(t))_{ij} = \mathrm{P}[N(t) = k, \varphi(t) = j | \varphi(0) = i], \qquad \text{for } 1 \leq i, j \leq m.$$

One easily verifies that $M_0(t) = e^{-\nu t}I$ and one proves by induction that

$$M_k(t) = e^{-\nu t}(\nu P)^k t^k / k! = e^{-\nu t}B^k t^k / k!, \tag{23}$$

for $k \geq 1$. Equation (23) holds for $k = 0$ and we assume that it holds for some $k - 1$. Conditioning on the epoch u of the first Poisson event, we find

13

that

$$M_k(t) = \int_0^t e^{-\nu u} \nu P M_{k-1}(t-u) \, du$$

$$= \int_0^t e^{-\nu u} B e^{-\nu(t-u)} B^{k-1}(t-u)^{k-1}/(k-1)! \, du$$

$$= e^{-\nu t} B^k/(k-1)! \int_0^t (t-u)^{k-1} \, du$$

$$= e^{-\nu t} B^k t^k/k!.$$

Taking $t = 1$, we find that

$$P[N(1) = k, T > 1] = \sum_{1 \le i \le M} \sum_{1 \le j \le M} \alpha_i P[N(t) = k, \varphi(t) = j | \varphi(0) = i]$$

$$= \boldsymbol{\alpha} e^{-\nu} B^k/k! \mathbf{1},$$

so that $P[T > 1] = \boldsymbol{\alpha} e^{-\nu} e^B \mathbf{1}$, and

$$P[N(1) = k | T > 1] = (\boldsymbol{\alpha} e^{-\nu} e^B \mathbf{1})^{-1} \boldsymbol{\alpha} e^{-\nu} B^k/k! \mathbf{1}$$

$$= (\boldsymbol{\alpha} e^B \mathbf{1})^{-1} \boldsymbol{\alpha} B^k/k! \mathbf{1}.$$

If $\boldsymbol{\alpha} = c\boldsymbol{\beta}$ for any scalar $c \le (\boldsymbol{\beta} \mathbf{1})^{-1}$, then $P[N(1) = k | T > 1] = p_k$ for all k. This concludes the proof. $\square$

Remark 4.5 If P is stochastic, then the random variable K has an unusual PH distribution, as it is either equal to zero or to infinity. Still, the argument in the proof of Theorem 4.4 holds true. Note that if P is stochastic, then $B\mathbf{1} = \nu\mathbf{1}$ and the distribution (18) is Poisson with parameter ν.

Example 4.6 This is a PH-Poisson distribution chosen to illustrate the combined effect of the conditional distribution imposed on the number of transitions among the phases. The representation is $(\boldsymbol{\beta}, B)$ with

$$B = \begin{bmatrix} 5 & .05 & 0 & 0 & 0 \\ .05 & 9 & .05 & 0 & 0 \\ 0 & .05 & 13 & .05 & 0 \\ 0 & 0 & .05 & 17 & .05 \\ 0 & 0 & 0 & .05 & 21 \end{bmatrix} \tag{24}$$

and

$$\boldsymbol{\beta} = \gamma \begin{bmatrix} 5. & 2.5 & 3. & 2.25 & 6. \end{bmatrix} \exp\{-\mathrm{diag}(5, 9, 13, 17, 21)\} \tag{25}$$
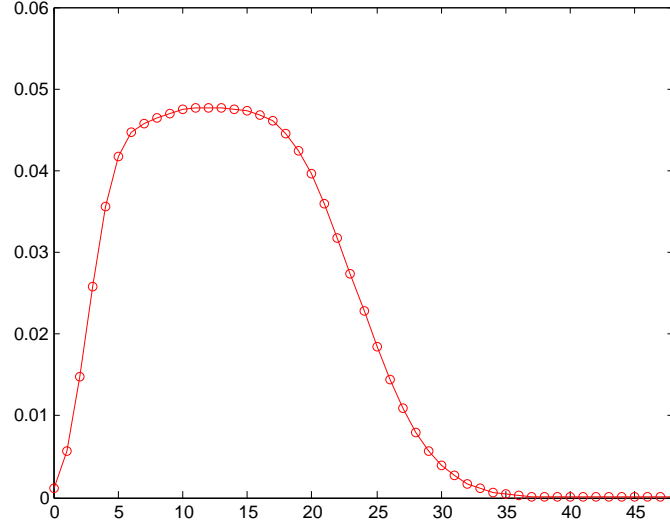
Figure 3: Density function for a PH-Poisson distribution with representation $(\boldsymbol{\beta}, B)$ given in (24, 25).

where the scaling factor γ is such that $\boldsymbol{\beta}e^B\mathbf{1} = 1$. Its first two moments are $\mu = 13.84$ and $\sigma^2 = 47.31$, and its density is given on Figure 3. The phase-type representation is $(\nu; \boldsymbol{\alpha}, P)$ with $\nu = 21.05$,

$$
P = \begin{bmatrix}
0.2375 & 0.0024 & 0 & 0 & 0 \\
0.0024 & 0.4276 & 0.0024 & 0 & 0 \\
0 & 0.0024 & 0.6176 & 0.0024 & 0 \\
0 & 0 & 0.0024 & 0.8076 & 0.0024 \\
0 & 0 & 0 & 0.0024 & 0.9976
\end{bmatrix}
\tag{26}
$$

and

$$
\boldsymbol{\alpha} \approx \begin{bmatrix} 0.99 & 0.91\,10^{-2} & 0.20\,10^{-3} & 0.27\,10^{-5} & 0.13\,10^{-6} \end{bmatrix},
$$

which is the vector $\boldsymbol{\beta}$ normalized so that $\boldsymbol{\alpha}\mathbf{1} = 1$.

Denote by $N^* = \sup\{k : \theta_k < 1\}$ the total number of Poisson events in $(0, 1)$. If $T \leq 1$, then $N(1) = K - 1 < N^*$, if $T > 1$, then $N(1) = N^* < K$. On the average, the Poisson process produces ν events in the interval $(0, 1)$, and the Poisson distribution has a relatively small standard deviation, so one expects N^* to take values close to $\nu \approx 21$. The matrix P in (26) is irreducible, albeit with a small probability of migration from one phase to another, so that the initial phase plays a significant role in the distribution of K.

If the initial phase is 1, then the absorption probability is about 0.76, and it is likely that K will be small; it is therefore necessary, for the condition $T > 1$ to be fulfilled, that the Poisson process produces few events in $(0, 1)$.

15

On the other hand, if the initial phase is 5, then the PH Markov chain will remain in that phase for a large number of transitions, it is likely that K will be large, so that T is likely to be much larger than 1, and it is not expected that the condition $[T > 1]$ puts much constraint on N^*.

5 EM algorithm

In this section, we exploit the probabilistic interpretation of PH-Poisson distributions given in Section 4.2, and we develop an EM algorithm for fitting PH-Poisson distributions into data samples.

The EM algorithm is a popular iterative method in statistics for computing maximum-likelihood estimates from data that is considered incomplete. The procedure can be explained briefly as follows. Let $\boldsymbol{\theta} \in \Omega$ be the set of parameters to be estimated. We denote by $\mathcal{X}$ a random complete data sample and by $f(\mathcal{X} \,|\, \boldsymbol{\theta})$ its conditional density function, given the parameters $\boldsymbol{\theta}$. The maximum-likelihood estimator $\hat{\boldsymbol{\theta}}$ is defined as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Omega} \log f(\mathcal{X} \,|\, \boldsymbol{\theta}).$$

For one reason or another, instead of observing the complete data sample $\mathcal{X}$, we observe an incomplete data sample $\mathcal{Y}$. Thus, $\mathcal{X}$ can be replaced by its sufficient statistic $(\mathcal{Y}, \mathcal{Z})$, where $\mathcal{Z}$ is the sufficient statistic of the unobserved data. As $\mathcal{X}$ is unobservable, instead of maximizing $\log f(\mathcal{X} \,|\, \boldsymbol{\theta})$ we maximize its conditional expectation given the incomplete data sample $\mathcal{Y} = \boldsymbol{y}$ and the current estimates $\boldsymbol{\theta}^{(s)}$, at each $(s+1)$th iteration for $s \geq 0$.

The EM algorithm can thus be decomposed into two steps:

- E-step—computing the conditional expectation of $\log f(\mathcal{X} \,|\, \boldsymbol{\theta})$ given the incomplete data sample $\boldsymbol{y}$ and the current estimates $\boldsymbol{\theta}^{(s)}$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)}) = \mathrm{E}[\log f(\mathcal{X} \,|\, \boldsymbol{\theta}) \,|\, \boldsymbol{y}, \boldsymbol{\theta}^{(s)}],$$

- M-step—obtaining the next set $\boldsymbol{\theta}^{(s+1)}$ of estimates by maximizing the expected log-likelihood determined in the E-step

$$\boldsymbol{\theta}^{(s+1)} = \arg\max_{\boldsymbol{\theta} \in \Omega} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s)}).$$

When fitting a PH-Poisson distribution into a data sample, the parameters to be estimated are $\boldsymbol{\theta} = \{\nu, \boldsymbol{\alpha}, P\}$. Without loss of generality, we assume that $\boldsymbol{\alpha}\mathbf{1} = 1$ in the chosen representation. By Theorem 4.4, an observation

16

y can be thought of as the number of Poisson events in the time interval $[0, 1]$, given that the transient Markov chain with the transition matrix P has not been absorbed at time $t = 1$. This observation can be considered incomplete as it tells us neither the initial phase $\varphi(0)$ of the Markov chain nor how it has evolved during $[0, 1]$; a *complete* observation can be represented by $x = (\varphi_0, \varphi_1, \ldots, \varphi_y)$, where φ_i is the phase of the Markov chain at the ith Poisson event and $\varphi_i \neq 0$ for all $i = 0, \ldots, y$. The conditional density of the complete observation x given $\boldsymbol{\theta}$ is

$$f(x \mid \boldsymbol{\theta}) = (\boldsymbol{\alpha} e^{\nu P} \mathbf{1})^{-1} \frac{\nu^y}{y!} \, \alpha_{\varphi_0} \prod_{i=0}^{y-1} p_{\varphi_i \varphi_{i+1}}.$$

Suppose that the complete data sample $\boldsymbol{x}$ contains n observations, each of which is denoted by $x^{[k]}$ and includes an incomplete observation $y^{[k]}$, for $k = 1, \ldots, n$. Then, the conditional density of $\boldsymbol{x}$ given $\boldsymbol{\theta}$ is

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = (\boldsymbol{\alpha} e^{\nu P} \mathbf{1})^{-n} \prod_{k=1}^{n} \frac{\nu^{y^{[k]}}}{y^{[k]}!} \prod_{k=1}^{n} \alpha_{\varphi_0^{[k]}} \prod_{k=1}^{n} \left(\prod_{i=0}^{y^{[k]}-1} p_{\varphi_i^{[k]} \varphi_{i+1}^{[k]}} \right)$$

$$= (\boldsymbol{\alpha} e^{\nu P} \mathbf{1})^{-n} \prod_{k=1}^{n} \frac{\nu^{y^{[k]}}}{y^{[k]}!} \prod_{i=1}^{m} \alpha_i^{S_i} \prod_{i=1}^{m} \prod_{j=1}^{m} p_{ij}^{N_{ij}},$$

where

$$S_i = \sum_{k=1}^{n} \mathbb{1}_{\{\varphi_0^{[k]}=i\}} \quad \text{for } i = 1, \ldots, m$$

is the number of complete observations in $\boldsymbol{x}$ with initial phase i, and

$$N_{ij} = \sum_{k=1}^{n} \sum_{t \geq 1} \mathbb{1}_{\{\varphi_{t-1}^{[k]}=i, \varphi_t^{[k]}=j\}} \quad \text{for } i, j = 1, \ldots, m$$

is the total number of jumps in $\boldsymbol{x}$ from phase i to phase j. Thus, the log-likelihood function is given by

$$\log f(\boldsymbol{x} \mid \boldsymbol{\theta}) = -n \log(\boldsymbol{\alpha} e^{\nu P} \mathbf{1}) + \sum_{k=1}^{n} y^{[k]} \log \nu - \sum_{k=1}^{n} \log(y^{[k]}!)$$

$$+ \sum_{i=1}^{m} S_i \log \alpha_i + \sum_{i=1}^{m} \sum_{j=1}^{m} N_{ij} \log p_{ij}. \tag{27}$$

17

Maximum-likelihood estimators To obtain closed-form expressions for the maximum-likelihood estimators $\hat{\boldsymbol{\theta}}$ is not straightforward. Applying the Karush-Kuhn-Tucker approach (see Chapter 12 in Nocedal and Wright (2000)), it can be verified that the maximization problem

$$\max_{\boldsymbol{\theta}} \; \log f(\boldsymbol{x} \mid \boldsymbol{\theta})$$

subject to

$$\boldsymbol{\alpha}\mathbf{1} = 1, \qquad P\mathbf{1} \leq \mathbf{1}, \qquad \nu > 0, \qquad p_{ij}, \alpha_i \geq 0 \quad \text{for } i,j = 1, \dots, m,$$

has the associated Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \lambda, \boldsymbol{\mu}) = \log f(\boldsymbol{x} \mid \boldsymbol{\theta}) - \lambda h(\boldsymbol{\theta}) - \sum_{i=1}^{2m+1} \mu_i g_i(\boldsymbol{\theta}),$$

where

$$\begin{aligned} h(\boldsymbol{\theta}) &= \boldsymbol{\alpha}\mathbf{1} - 1, \\ g_i(\boldsymbol{\theta}) &= 1 - \sum_{j=1}^{m} p_{ij} \quad \text{for } i = 1, \dots, m, \\ &= \alpha_{i-m} \qquad \text{for } i = m+1, \dots, 2m, \\ &= \nu \qquad\qquad \text{for } i = 2m+1, \end{aligned}$$

and λ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{2m+1}) \geq \mathbf{0}$ denote the Lagrangian multipliers associated with the equality constraint $h(\boldsymbol{\theta}) = 0$, the inequality constraints $g_i(\boldsymbol{\theta}) \geq 0$ for $i = 1, \dots, 2m$ and $g_{2m+1}(\boldsymbol{\theta}) > 0$, respectively. The KKT conditions, which are first-order necessary conditions for constrained optimization problems, imply that the maximum-likelihood estimators $\hat{\boldsymbol{\theta}} = (\hat{\nu}, \hat{\boldsymbol{\alpha}}, \hat{P})$ must satisfy the following constraints

$$\hat{\boldsymbol{\alpha}} e^{\hat{\nu}\hat{P}} (\hat{\nu}\hat{P}\mathbf{1} - \frac{\sum_{k=1}^{n} y^{[k]}}{n}\mathbf{1}) = 0 \tag{28}$$

$$\hat{\alpha}_i = \frac{S_i}{\hat{\eta}_i} (\sum_{j=1}^{m} \frac{S_j}{\hat{\eta}_j})^{-1} \quad \text{for } i = 1, \dots, m, \tag{29}$$

$$\frac{n\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}} e^{\hat{\nu}\hat{P}}\mathbf{1}} \int_0^{\hat{\nu}} e^{(\hat{\nu}-u)\hat{P}} \boldsymbol{e}_i \boldsymbol{e}_j^{\mathrm{T}} e^{u\hat{P}} \, du\mathbf{1} - \frac{N_{ij}}{\hat{p}_{ij}} \leq 0, \tag{30}$$

for $i, j = 1, \dots, m$, where $\hat{\eta}_i = \boldsymbol{e}_i^{\mathrm{T}} e^{\hat{\nu}\hat{P}}\mathbf{1}$ and $\boldsymbol{e}_i$ is the column vector of size m with the ith component being 1 and all other components being 0.

Recall from Remark 4.5, that if P is stochastic then the PH-Poisson distribution with representation $(\nu, \boldsymbol{\alpha}, P)$ is a Poisson distribution with parameter ν. In this case, the constraints (28)–(30) simplify considerably: the first

18

implies that $\hat{\nu} = \sum_{k=1}^{n} y^{[k]}/n$, the well-known maximum-likelihood estimator for the parameter of a Poisson distribution; the second becomes $\hat{\alpha}_i = S_i/n$, the maximum-likelihood estimator for the initial vector of a discrete PH distribution (see Asmussen et al. (1996)); and the third reduces to

$$n\hat{\boldsymbol{\alpha}} \int_0^{\hat{\nu}} e^{x(\hat{P}-I)} \, \mathrm{d}x \boldsymbol{e}_i - \frac{N_{ij}}{\hat{p}_{ij}} \leq 0,$$

or, equivalently,

$$\hat{\nu}\hat{p}_{ij}\hat{\boldsymbol{\alpha}} \int_0^1 e^{\hat{\nu}(\hat{P}-I)x} \, \mathrm{d}x \boldsymbol{e}_i - \frac{N_{ij}}{n} \leq 0, \tag{31}$$

for $i, j = 1, \ldots, m$. As $\hat{P}$ is stochastic, summing the left-hand side of (31) over i and j gives us

$$\hat{\nu}\hat{\boldsymbol{\alpha}} \int_0^1 e^{\hat{\nu}(\hat{P}-I)x} \, \mathrm{d}x \mathbf{1} - 1/n \sum_{i,j=1}^m N_{ij} = \hat{\nu} - 1/n \sum_{i,j=1}^m N_{ij} = 0,$$

which implies that (31) is an equality for all $i, j = 1, \ldots, m$.

Conditional expectation Thanks to the linear nature of $\log f(\mathcal{X} \mid \boldsymbol{\theta})$ in the unobserved data $\mathcal{Z} = \{S_i, N_{ij} : i, j = 1, \ldots, m\}$, the computation of the conditional expectation of $\log f(\mathcal{X} \mid \boldsymbol{\theta}^{(s)})$ at the $(s+1)$th iteration reduces to the computation of $\mathrm{E}[\mathcal{Z} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(s)}]$:

$$
\begin{aligned}
\mathrm{E}[S_i \mid \boldsymbol{y}, \boldsymbol{\theta}^{(s)}] &= \sum_{k=1}^n \mathrm{E}[\mathbb{1}_{\{\varphi_0^{[k]}=i\}} \mid y^{[k]}, \boldsymbol{\theta}^{(s)}] \\
&= \sum_{k=1}^n \frac{\mathrm{P}[\varphi_0^{[k]} = i \mid \boldsymbol{\theta}^{(s)}] \, \mathrm{P}[Y^{[k]} = y^{[k]} \mid \boldsymbol{\theta}^{(s)}, \varphi_0^{[k]} = i]}{\mathrm{P}[Y^{[k]} = y^{[k]} \mid \boldsymbol{\theta}^{(s)}]} \\
&= \sum_{k=1}^n \frac{\alpha_i^{(s)} \, \boldsymbol{e}_i^{\mathrm{T}} \, (P^{(s)})^{y^{[k]}} \mathbf{1}}{\boldsymbol{\alpha}^{(s)} \, (P^{(s)})^{y^{[k]}} \mathbf{1}} \quad \text{for } i = 1, \ldots, m, \tag{32}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{E}[N_{ij} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(s)}] &= \sum_{k=1}^n \sum_{t=1}^{y^{[k]}} \mathrm{E}[\mathbb{1}_{\{\varphi_{t-1}^{[k]}=i, \varphi_t^{[k]}=j\}} \mid y^{[k]}, \boldsymbol{\theta}^{(s)}] \\
&= \sum_{k=1}^n \sum_{t=1}^{y^{[k]}} \frac{\mathrm{P}[\varphi_{t-1}^{[k]} = i \mid \boldsymbol{\theta}^{(s)}] \, \mathrm{P}[\varphi_t^{[k]} = j \mid \boldsymbol{\theta}^{(s)}, \varphi_{t-1}^{[k]} = i] \, \mathrm{P}[Y^{[k]} = y^{[k]} \mid \boldsymbol{\theta}^{(s)}, \varphi_t^{[k]} = j]}{\mathrm{P}[Y^{[k]} = y^{[k]} \mid \boldsymbol{\theta}^{(s)}]}
\end{aligned}
$$

$$= \sum_{k=1}^{n} \frac{\boldsymbol{\alpha}^{(s)} \sum_{t=1}^{y^{[k]}} (P^{(s)})^{t-1} \, \boldsymbol{e}_i \, p_{ij}^{(s)} \, \boldsymbol{e}_j^{\mathrm{T}} (P^{(s)})^{y^{[k]}-t} \, \mathbf{1}}{\boldsymbol{\alpha}^{(s)} (P^{(s)})^{y^{[k]}} \mathbf{1}} \quad \text{for } i,j = 1,\dots,m. \quad (33)$$

New estimates In the M-step, we obtain the new estimates $\boldsymbol{\theta}^{(s+1)} = (\nu^{(s+1)}, \boldsymbol{\alpha}^{(s+1)}, P^{(s+1)})$ by maximizing the log-likelihood (27) where $\{S_i, N_{ij} : i,j = 1,\dots,p\}$ are replaced by their conditional expectations $\mathrm{E}[S_i \,|\, \boldsymbol{y}, \boldsymbol{\theta}^{(s)}]$ and $\mathrm{E}[N_{ij} \,|\, \boldsymbol{y}, \boldsymbol{\theta}^{(s)}]$ evaluated in the E-step. The maximization problem to be solved in this step is as follows

$$\max_{\boldsymbol{\theta}} \ \log f(\boldsymbol{y}, \mathrm{E}[S_i \,|\, \boldsymbol{\theta}^{(s)}], \mathrm{E}[N_{ij} \,|\, \boldsymbol{\theta}^{(s)}] \,|\, \boldsymbol{\theta})$$

subject to

$$\boldsymbol{\alpha}\mathbf{1} = 1, \qquad P\mathbf{1} \leq \mathbf{1}, \qquad \nu > 0, \qquad p_{ij}, \alpha_i \geq 0 \quad \text{for } i,j = 1,\dots,m.$$

We implemented the EM algorithm in MATLAB and experimented with samples simulated from different PH-Poisson distributions. Below are the results of one such experiment.

Example 5.1 We used the PH-Poisson distribution $(\nu; \boldsymbol{\alpha}, P)$ given in Example 4.6 to generate a sample with 1500 observations. The chosen initial parameters are $\nu^{(0)} = 10$, $\boldsymbol{\alpha}^{(0)} = \begin{bmatrix} 0.1 & 0.2 & 0.4 & 0.2 & 0.1 \end{bmatrix}$ and

$$P^{(0)} = \mathrm{diag}(0.5, 0.3, 0.5, 0.7, 0.1).$$

The estimated parameters obtained after 25 iterations of the EM algorithm are $\nu^{(25)} = 20.4290$, $\boldsymbol{\alpha}^{(25)} = \begin{bmatrix} 0.9054 & 0.060 & 0.0335 & 0.0000 & 0.0010 \end{bmatrix}$, and

$$P^{(25)} = \begin{bmatrix} 0.2401 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.2610 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.4543 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.9088 & 0.0000 \\ 0.9939 & 0.0026 & 0.0000 & 0.0000 & 0.0035 \end{bmatrix}.$$

The Manhattan norm $||\cdot||_1$ of the difference between the true density and the empirical data is 0.1109, between the true density and the estimated density is 0.1043, and between the empirical data and the estimated density is 0.1400. We plot four densities in Figure 4: that for the true PH-Poisson distribution, the empirical data, the initial density and the estimated density.
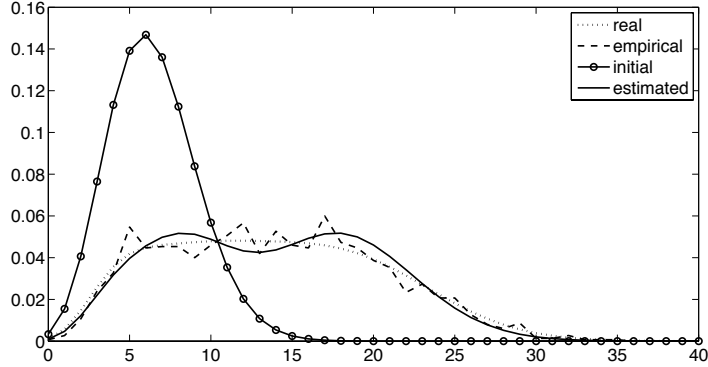
Figure 4: Density function for the true PH-Poisson distribution (the dotted curve), empirical data (the dashed curve), the initial density (the curve marked with ∘) and the estimated density (the continuous curve).

It is well-known that although the sequence $\{\boldsymbol{\theta}^{(s)}\}_{s \geq 1}$ computed with the EM algorithm always converges, it does not always converge to the maximum-likelihood estimator $\hat{\boldsymbol{\theta}}$, but possibly to some local maximum or stationary value of $\log f(\mathcal{X}|\boldsymbol{\theta})$. The warranty of global convergence for the EM algorithm depends on properties of the conditional density of the incomplete data $\mathcal{Y}$ given $\boldsymbol{\theta}$, and sometimes also on the starting point $\boldsymbol{\theta}^{(0)}$. We refer to Dempster et al. (1977) for further details on the EM algorithm, and to Wu (1983) for its convergence properties.

Our experiments were performed using the MATLAB optimization routine fmincon to solve the maximization problem in the M-step. They indicated that the results were highly sensitive to the choice of $\boldsymbol{\theta}^{(0)}$. When the starting point was chosen randomly, we observed that the EM algorithm often converged to a Poisson distribution with parameter $\sum_{k=1}^{n} y^{[k]}/n$, even if this was a rather poor fit for the given sample. Convergence to a good fit was obtained with a starting point that either shares the same structure of zeros with the true parameters $\boldsymbol{\alpha}$ and P, or has a strictly positive $\boldsymbol{\alpha}^{(0)}$ and a diagonal matrix $P^{(0)}$—a mixture of Poisson distributions.

The latter choice is obviously more practical when the structure of the true parameters is not known a priori. Empirically, a diagonal $P^{(0)}$ proved to be a good starting point even if the true matrix P is not diagonal. Note that, unlike its counterpart for fitting discrete Phase-type distributions in Asmussen et al. (1996), the EM algorithm for fitting PH-Poisson distributions does not necessarily preserve the initial structure. This is due to the term $-n \log(\boldsymbol{\alpha}e^{\nu P}\mathbf{1})$ in (27). Consequently, when starting with a diagonal $P^{(0)}$ the EM algorithm does not necessarily converge to a diagonal P. An

interesting question for future research is to explain why mixtures of Poisson distributions serve as good starting points in the EM algorithm for fitting Phase-type Poisson distributions.

Acknowledgment

References

[1] S. Asmussen, O. Nerman, and M. Olsson. Fitting Phase-Type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser.*, 39:1–38, 1977.

[3] K.-T. Eisele. Recursions for compound phase distributions. *Insurance: Mathematics and Economics*, 38:149–156, 2006.

[4] D. E. Knuth. *The Art of Computer Programming: Fundamental Algorithms*, volume 1. Addison–Wesley, Reading, MA, 1968.

[5] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia PA, 1999.

[6] M. F. Neuts. Probability distributions of phase type. In *Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206. University of Louvain, Belgium, 1975.

[7] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University Press, Baltimore, MD, 1981.

[8] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2000.

[9] H. H. Panjer. Recursive evaluation of a family of compound distributions. *ASTIN Bulletin*, 12:22–26, 1981.

[10] J. Ren. Recursive formulas for compound phase distributions — univariate and bivariate cases. *ASTIN Bulletin*, 40:615–629, 2010.

[11] K. K. Siaw, X. Wu, D. Pitt, and Y. Wang. Matrix-form recursive evaluation of the aggregate claims distribution revisited. *Annals of Actuarial Science*, 5:163–179, 2011.

[12] B. Sundt and W. S. Jewell. Further results on recursive evaluation of compound distributions. *ASTIN Bulletin*, 12:27–39, 1981.

[13] M. Telek. The minimal coefficient of variation of discrete phase-type distributions. In G. Latouche and P. Taylor, editors, *Advances in Algorithmic Methods for Stochastic Models – Proceedings of the 3rd International Conference on Matrix-Analytic Methods*, pages 391–400. Notable Publications Inc, NJ, 2000.

[14] H. S. Wilf. The asymptotic behavior of the Stirling numbers of the first kind. *J. Comb. Theory, Ser. A*, 64(2):344–349, 1993.

[15] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

[16] X. Wu and S. Li. Matrix-form recursions for a family of compound distributions. *ASTIN Bulletin*, 40:351–368, 2010.